(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 December 2001 (27.12.2001)

PCT

(10) International Publication Number
WO 01/99043 A1

(51) International Patent Classification[7]: G06K 9/62

(21) International Application Number: PCT/US01/19376

(22) International Filing Date: 19 June 2001 (19.06.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/212,404 19 June 2000 (19.06.2000) US

(71) Applicant: **CORRELOGIC SYSTEMS, INC.** [US/US]; Suite 300, 6701 Democracy Boulevard, Behtesda, MD 20817 (US).

(72) Inventor: **HITT, Ben**; 1910 Curie Drive, Severn, MD 21144 (US).

(74) Agents: **GLOVER, Gregory, J.** et al.; Ropes & Gray, Suite 800, 1301 K Street, N.W., Washington, DC 20005-3333 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/99043 A1

(54) Title: HEURISTIC METHOD OF CLASSIFICATION

(57) Abstract: The invention concerns heuristic algorithms for the classification of Objects. A first learning algorithm comprises a genetic algorithm that is used to abstract a data stream associated with each Object and a pattern recognition algorithm that is used to classify the Objects and measure the fitness of the chromosomes of the genetic algorithm. The learning algorithm is applied to a training data set. The learning algorithm generates a classifying algorithm, which is used to classify or categorize unknown Objects. The invention is useful in the areas of classifying texts and medical samples, predicting the behavior of one financial market based on price changes in others and in monitoring the state of complex process facilities to detect impending failures.

Heuristic Method of Classification

This application claims benefit under 35 U.S.C. sec. 119(e)(1) of the priority of application Serial No. 60/212,404, filed June 19, 2000, which is hereby incorporated by reference in its entirety.

I. Field of the Invention

The field of the invention concerns a method of analyzing and classifying objects which can be represented as character strings, such as documents, or strings or tables of numerical data, such as changes in stock market prices, the levels of expression of different genes in cells of a tissue detected by hybridization of mRNA to a gene chip, or the amounts of different proteins in a sample detected by mass spectroscopy. More specifically, the invention concerns a general method whereby a classification algorithm is generated and verified from a learning data set consisting of pre-classified examples of the class of objects that are to be classified. The pre-classified examples having been classified by reading in the case of documents, historical experience in the case of market data, or pathological examination in the case of biological data. The classification algorithm can then be used to classify previously unclassified examples. Such algorithms are generically termed data mining techniques. The more commonly applied data mining techniques, such as multivariate linear regression and non linear feed-forward neural networks have an intrinsic shortcoming, in that, once developed, they are static and cannot recognize novel events in a data stream. The end result is that novel events often get misclassified. The invention concerns a solution to this shortcoming through an adaptive mechanism that can recognize novel events in a data stream.

II. Background of the Invention

The invention uses genetic algorithms and self organizing adaptive pattern recognition algorithms. Genetic algorithms were described initially by Professor John H. Holland. (J.H. Holland, *Adaptation in Natural and Artificial Systems,* MIT Press 1992, see also U.S. patent No. 4,697,242 and No. 4,881,178). A use of a genetic algorithm for pattern recognition is described in U.S. patent No. 5,136,686 to Koza, see column 87.

Self organizing pattern recognition has been described by Kohonen. (T. Kohonen, *Self Organizing and Associative Memory, 8 Series in Information Sciences,* Springer Verlag, 1984; Kohonen, T, *Self-organizing Maps,* Springer Verlag, Heidelberg 1997). The use of self organizing maps in adaptive pattern recognition was described by Dr. Richard Lippman of the Massachusetts Institute of Technology.

III. Summary of the Invention

The invention consists of two related heuristic algorithms, a classifying algorithm and a learning algorithm, which are used to implement classifying methods and learning methods. The parameters of the classifying algorithm are determined by the application of the learning algorithm to a training or learning data set. The training data set is a data set in which each item has already been classified. Although the following method is described without reference to digital computers, it will be understood by those skilled in the art that the invention is intended for implementation as computer software. Any general purpose computer can be used; the calculations according to the method are not unduly extensive. While computers having parallel processing facility could be used for the invention, such processing capabilities are not necessary for the practical use of the learning algorithm of the invention. The classifying algorithm requires only a minimal amount of computation.

The classifying method of the invention classifies Objects according to a data stream that is associated with the Object. Each Object in the invention is characterized by a data stream, which is a large number, at least about 100 data points, and can be 10,000 or more data points. A data stream is generated in a way that allows for the individual datum in data streams of different samples of the same type of Object to be correlated one with the other.

Examples of Objects include texts, points in time in the context of predicting the direction of financial markets or the behavior of a complex processing facility, and biological samples for medical diagnosis. The associated data streams of these Objects are the distribution of trigrams in the text, the daily changes in price of publicly traded stocks or commodities, the instantaneous readings of a number of pressure, temperature and flow readings in the processing facility such as an oil refinery, and a mass spectrum of some subset of the proteins found in the sample, or the intensity mRNA hybridization to an array of different test polynucleotides.

Thus, generically the invention can be used whenever it is desired to classify Objects into one of several categories, *e.g.*, which typically is two or three categories, and the Objects are associated with extensive amounts of data, *e.g.*, typically thousands of data points. The term "Objects" is capitalized herein to indicate that Objects has a special meaning herein in that it refers collectively to tangible objects, *e.g.*, specific samples, and intangible objects, *e.g.*, writings or texts, and totally abstract objects, *e.g.*, the moment in time prior to an untoward event in a complex processing facility or the movement in the price of a foreign currency.

The first step of the classifying method is to calculate an Object vector, *i.e.*, an ordered set of a small number of data points or scalers (between 4 and 100, more typically between 5 and 30) that is derived from the data stream associated with the

Object to be classified. The transformation of the data steam into an Object vector is termed "abstraction." The most simple abstraction process is to select a number of points of the data stream. However, in principle the abstraction process can be performed on any function of the data stream. In the embodiments presented below abstraction is performed by selection of a small number of specific intensities from the data stream.

In one embodiment, the second step of the classifying method is to determine in which data cluster, if any, the vector rests. Data clusters are mathematical constructs that are the multidimensional equivalents of non-overlapping "hyperspheres" of fixed size in the vector space. The location and associated classification or "status" of each data cluster is determined by the learning algorithm from the training data set. The extent or size of each data cluster and the number of dimensions of the vector space is set as a matter of routine experimentation by the operator prior to the operation of the learning algorithm. If the vector lies within a known data cluster, the Object is given the classification associated with that cluster. In the most simple embodiments the number of dimensions of the vector space is equal to the number of data points that is selected in the abstraction process. Alternatively, however, each scaler of the Object vector can be calculated using multiple data points of the data stream. If the Object vector rests outside of any known cluster, a classification can be made of atypia, or atypical sample.

In an alternative embodiment, the definition of each data cluster as a hypersphere is discarded and the second step is performed by calculating the match parameter $\rho = \Sigma (\min (|I_i|, |W_i|) / \Sigma (|W_i|)$, where I_i are the scalers of the Object vector and W_i are the scalers of the centroid of the preformed classifying vector. The match parameter ρ is also termed a normalized "fuzzy" AND. The Object is then classified

according to the classification of the preformed vector to which it is most similar by this metric. The match parameter is 1 when the Object vector and the preformed vector are identical and less than 1 in all other cases.

The learning algorithm determines both the details of abstraction process and the identity of the data clusters by utilizing a combination of known mathematical techniques and two pre-set parameters. A user pre-sets the number of dimensions of the vector space and the size of the data clusters or, alternatively, the minimum acceptable level of the "fuzzy AND" match parameter ρ. As used herein the term "data cluster" refers to both a hypersphere using a Euclidean metric and preformed classified vectors using a "fuzzy AND" metric.

Typically the vector space in which the data clusters lie is a normalized vector space so that the variation of intensities in each dimension is constant. So expressed the size of the data cluster using a Euclidean metric can be expressed as minimum percent similarity among the vectors resting within the cluster.

In one embodiment the learning algorithm can be implemented by combining two different types of publicly available generic software, which have been developed by others and are well known in the field: (1) a genetic algorithm (J.H. Holland, *Adaptation in Natural and Artificial Systems,* MIT Press 1992) that processes a set of logical chromosomes[1] to identify an optimal logical chromosome that controls the abstraction of the data steam and (2) an adaptive self-organizing pattern recognition system (see, T. Kohonen, *Self Organizing and Associative Memory, 8 Series in*

[1] The term logical chromosome is used in connection with genetic learning algorithms because the logical operations of the algorithm are analogous to reproduction, selection, recombination and mutation. There is, of course, no biological embodiment of a logical chromosome in DNA or otherwise. The genetic learning algorithms of the invention are purely computational devices, and should not be confused with schemes for biologically-based information processing.

Information Sciences, Springer Verlag, 1984; Kohonen, T, *Self-organizing Maps,* Springer Verlag, Heidelberg 1997), available from Group One Software, Greenbelt, MD, which identifies a set of data clusters based on any set of vectors generated by a logical chromosome. Specifically the adaptive pattern recognition software maximizes the number of vectors that rest in homogeneous data clusters, *i.e.,* clusters that contain vectors of the learning set having only one classification type.

To use a genetic algorithm each logical chromosome must be assigned a "fitness." The fitness of each logical chromosome is determined by the number of vectors in the training data set that rest in homogeneous clusters of the optimal set of data clusters for that chromosome. Thus, the learning algorithm of the invention combines a genetic algorithm to identify an optimal logical chromosome and an adaptive pattern recognition algorithm to generate an optimal set of data clusters and a the fitness calculation based on the number of sample vectors resting in homogeneous clusters. In its broadest embodiment, the learning algorithm of the invention consists of the combination of a genetic algorithm, a pattern recognition algorithm and the use of a fitness function that measures the homogeneity of the output of the pattern recognition algorithm to control the genetic algorithm.

To avoid confusion, it should be noted that the number of data clusters is much greater than the number of categories. The classifying algorithms of the examples below sorted Objects into two categories, *e.g.,* documents into those of interest and those not of interest, or the clinical samples into benign or malignant. These classifying algorithms, however, utilize multiple data clusters to perform the classification. When the Object is a point in time, the classifying algorithm may utilize more than two categories. For example, when the invention is used as a predictor of foreign exchange rates, a tripartite scheme corresponding to rising, falling

and mixed outlooks would be appropriate. Again, such a tripartite classifying algorithm would be expected to have many more than three data clusters.

IV. Detailed Description of the Invention

In order to practice the invention the routine practitioner must develop a classifying algorithm by employing the learning algorithm. As with any heuristic method, some routine experimentation is required. To employ the learning algorithm, the routine practitioner uses a training data set and must experimentally optimize two parameters, the number of dimensions and the data cluster size.

Although there is no absolute or inherent upper limit on the number of dimensions in the vector, the learning algorithm itself inherently limits the number of dimensions in each implementation. If the number of dimensions is too low or the size of the cluster is too large, the learning algorithm fails to generate any logical chromosomes that correctly classify all samples with an acceptable level of homogeneity. Conversely, the number of dimensions can be too large. Under this circumstance, the learning algorithm generates many logical chromosomes that have the maximum possible fitness early in the learning process and, accordingly, there is only abortive selection. Similarly, when the size of the data clusters is too small, the number of clusters will be found to approach the number of samples in the training data set and, again, the routine practitioner will find that a large number of logical chromosomes will yield a set of completely homogeneous data clusters.

Although the foregoing provide general guidance for the selection of the number of dimensions and the data cluster size for a classifying algorithm, it should be understood that the true test of the value of a classifying algorithm is its ability to correctly classify data streams that are independent of the data stream in the training data set. Therefore, the routine practitioner will understand that a portion of the

learning data set must be reserved to verify that the classification algorithm is functioning with an error rate, that is acceptable for the intended purpose. The particular components of the invention are described in greater detail below.

A. The Data Stream and Types of Objects

The classification of Objects and the generation of the associated data stream depend upon the nature of the problem to be addressed. The general principles are illustrated by the following examples.

Documents: In one embodiment the invention provides a method for the computerized classification documents. For example, one may want to extract the documents of interest from a data base consisting of a number of documents too large to review individually. For these circumstances, the invention provides a computerized algorithm to identify a subset of the database most likely to contain the documents of interest. Each document is an Object, the data stream for each document consists of the histogram representing the frequency of each of the 17576 (26^3) three letter combinations (trigrams) found in the document after removal of spaces and punctuation. Alternatively, a histogram of the 9261 trigrams of consonants can be prepared after the further removal of vowels from the document. The training data set consists of a sample of the appropriate documents that have been classified as "of interest" or "not of interest," according to the needs of the user.

Financial Markets: It is self-evident that financial markets respond to external events and are interrelated to each other in a consistent fashion; for example, foreign exchange rates are influenced by the attractiveness of investment opportunities. However, the direction and extent of the response to an individual event can be difficult to predict. In one embodiment, the invention provides an algorithm computerized prediction of prices in one market based on the movement in

prices in another. Each point in time is an Object, for example hourly intervals, the data stream for hour consists of the histogram of the change in price of publicly traded securities in the major stock markets in the relevant countries, *e.g.*, the New York and London stock exchanges where the exchange rate of the pound and dollar are of interest. The training data set consists of the historical record such price changes that has been classified as preceding a rise or fall in the dollar:pound rate.

Processing Facilities: In a complex processing facility, such as an oil refinery, oil field or petrochemical plant, the pressure, temperature, flow and status of multiple valves and other controls (collectively the "status values") are constantly monitored and recorded. There is a need to detect impending untoward events before the untoward event becomes a catastrophic failure. The present invention provides a computerized algorithm to classify each point in time as either a high-risk or normal-risk time point. The data stream consists of the status values for each point in time. The training data set consists of the historical record of the status values classified as either preceding an untoward event or as preceding normal operation.

Medical Diagnosis: The invention can be used in the analysis of a tissue sample for medical diagnosis, *e.g.*,, for analysis of serum or plasma. The data stream can be any reproducible physical analysis of the tissue sample that results in 2,000 or more measurements that can be quantified to at least 1 part per thousand (three significant figures). Time of flight mass spectra of proteins are particularly suitable for the practice of the invention. More specifically, matrix assisted laser desorption ionization time of flight (MALDI-TOF) and surface enhanced laser desorption ionization time of flight (SELDI-TOF) spectroscopy. See generally WO 00/49410.

The data stream can also include measurements that are not inherently organized by a single ordered parameter such as molecular weight, but have an

arbitrary order. Thus, DNA microarray data that simultaneously measures the expression levels of 2,000 or more genes can be used as a data stream when the tissue sample is a biopsy specimen, recognizing that the order of the individual genes is the data stream is arbitrary.

Specific diseases where the present invention is particularly valuable occur when early diagnosis is important, but technically difficult because of the absence of symptoms and the disease may be expected to produce differences that are detectable in the serum because of the metabolic activity of the pathological tissue. The early diagnosis of malignancies are a primary focus of the use of the invention. The working example illustrates the diagnosis of prostatic carcinoma, similar trials for the diagnosis of ovarian cancers have been performed.

It should be noted that a single data stream from a patient sample can be analyzed for multiple diagnoses using the method of the invention. The additional cost of such multiple analysis would be trivial because the steps specific to each diagnosis are computational only.

B. The Abstraction Process and Logical Chromosome

The first step in the classifying process of the invention is the transformation or abstraction of the data stream into a characteristic vector. The data may be conveniently normalized prior to abstraction by assigning the overall peak a arbitrary value of 1.0 and all other points given fractional values. The most simple abstraction of a data stream consists of the selection of a small number of data points. Those skilled in the will recognize that more complex functions of multiple points could be constructed such as averages over intervals or more complex sums or differences between data points that are at predetermined distance from a selected prototype data point. Such functions of the intensity values of the data stream could also be used and

'D

are expected to function equivalently to the simple abstract illustrated in the working examples.

The skilled will also appreciate that routine experimentation can determine whether abstraction by taking the instantaneous slope at arbitrary points could also function in the present invention. Accordingly, such routinely available variations of the illustrated working examples are within the scope of the invention.

A feature of the invention is the use of a genetic algorithm to determine the data points which are used to calculate the characteristic vector. In keeping with the nomenclature of the art, the list of the specific points to be selected is termed a logical chromosome. The logical chromosomes contain as many "genes" as there are dimensions of the characteristic vector. Any set of the appropriate number of data points can be a logical chromosome, provided only that no gene of a chromosome is duplicated. The order of the genes has no significance to the invention.

Those skilled in the art appreciate that a genetic algorithm can be used when two conditions are met. A particular solution to a problem must be able to be expressed by a set or string of fixed size of discrete elements, which elements can be numbers or characters, and the strings can be recombined to yield further solutions. One must also be able to calculate a numerical value of the relative merit of each solution, its fitness. Under these circumstances the details of the genetic algorithm are unrelated to the problem whose solution is sought. Accordingly, for the present invention, generic genetic algorithm software may be employed. The algorithms in PGAPack libraries, available from Argonne National Laboratory is suitable. The calculation of the fitness of any particular logical chromosome is discussed below.

The first illustrative example concerns a corpus of 100 documents, which were randomly divided into a training set of 46 documents and a testing set of 54

11

documents. The documents consisted of State of the Union addresses, selections from the book *The Art of War* and articles from the Financial Times. The distribution of trigrams for each document was calculated. A vector space of 25 dimensions and a data cluster size in each dimension of 0.35 times the range of values in that dimension was selected. The genetic algorithms were initialized with about 1,500 randomly chosen logical chromosomes. As the algorithm progressed the more fit logical chromosomes are duplicated and the less fit are terminated. There is recombination between chromosomes and mutation, which occurs by the random replacement of an element of a chromosome. It is not an essential feature of the invention that the initially selected collection of logical chromosome be random. Certain prescreening of the total set of data streams to identify those data points having the highest variability may be useful, although such techniques may also introduce an unwanted initialization bias. Those skilled in the art appreciate that the initial set of chromosomes, the mutation rate and other boundary conditions for the genetic algorithm are not critical to its function.

C. The Pattern Recognition Process and Fitness Score Generation

The fitness score of each of the logical chromosomes that are generated by the genetic algorithm is calculated. The calculation of the fitness score requires an optimal set of data clusters be generated for each logical chromosome that is tested. Data clusters are simply the volumes in the vector space in which the Object vectors of the training data set rest. The method of generating the optimal set of data clusters is not critical to the invention and will be considered below. However, whatever method is used to generate the data cluster map, the map is constrained by the following rules: each data cluster should be located at the centroid of the data points

12

that lie within the data cluster, no two data clusters may overlap and the dimension of each cluster in the normalized vector space is fixed prior to the generation of the map.

The size of the data cluster is set by the user during the training process. Setting the size too large results in a failure find any chromosomes that can successfully classify the entire training set, conversely setting the size to low results in a set of optimal data clusters in which the number of clusters approaches the number of data points in the training set. More importantly, a too small setting of the size of the data cluster results in "overfitting," which is discussed below.

The method used to define the size of the data cluster is a part of the invention. The cluster size can be defined by the maximum of the equivalent of the Euclidean distance (root sum of the squares) between any two members of the data cluster. A data cluster size that corresponds to a requirement of 90% similarity is suitable for the invention when the data stream is generated by SELDI-TOF mass spectroscopy data. Somewhat large data clusters have been found useful for the classification of texts. Mathematically, 90% similarity is defined by requiring that the distance between any two members of a cluster is less than 0.1 of the maximum distance between two points in a normalized vector space. For this calculation, the vector space is normalized so that the range of each scalar of the vectors within the training data set is between 0.0 and 1.0. Thus normalized, the maximal possible distance between any two vectors in the vector space is then root N, where N is the number of dimensions. The Euclidean diameter of each cluster is then 0.1 x root(N).

The specific normalization of the vector space is not a critical feature of the method. The foregoing method was selected for ease of calculation. Alternative normalization can be accomplished by scaling each dimension not to the range but so

that each dimension has an equal variance. Non-Euclidean metrics, such as vector product metrics can be used.

Those skilled in the art will further recognize that the data stream may be converted into logarithmic form if the distribution of values within the data stream is log normal and not normally distributed.

Once the optimal set of data clusters for a logical chromosome has been generated, the fitness score for that chromosome can be calculated. For the present invention, the fitness score of the chromosome roughly corresponds to the number of vectors of the training data set that rest in clusters that are homogeneous, *i.e.*, clusters that contain the characteristic vectors from samples having a single classification. More precisely, the fitness score is calculated by assigning to each cluster a homogeneity score, which varies from 0.0 for homogeneous clusters to 0.5 for clusters that contain equal numbers of malignant and benign sample vectors. The fitness score of the chromosome is the average fitness score of the data clusters. Thus, a fitness score of 0.0 is the most fit. There is a bias towards logical chromosomes that generate more data clusters, in that when two logical chromosomes that have equal numbers of errors in assigning the data, the chromosome that generates the greater number of clusters will have a lower average homogeneity score and thus a better fitness score.

Publicly available software for generating using the self-organizing map is has been given several names, one is a "Lead Cluster Map" and can be implemented by generic software that is available as Model 1 from Group One Software (Greenbelt, MD).

An alternative embodiment of the invention utilizes a non-Euclidean metric to establish the boundaries of the data clusters. A metric refers to a method of

14

measuring distance in a vector space. The alternative metric for the invention can be based on a normalized "fuzzy AND" as defined above. Soft ware that implements an adaptive pattern recognition algorithm based on the "fuzzy AND" metric is available from Boston University under the name Fuzzy ARTMAP.

D. Description and Verification of Specific Embodiments

Those skilled in the art understand that the assignment of the entire training data set into homogeneous data clusters is not in itself evidence that the classifying algorithm is effectively operating at an acceptable level of accuracy. Thus, the value of the classifying algorithm generated by a learning algorithm must be tested by its ability to sort a set of data other than the training data set. When a learning algorithm generates a classifying algorithm that successfully assigns the training data set but only poorly assigns the test data set, the training data is said to be overfitted by learning algorithm. Overfitting results when the number of dimensions is too large and/or the size of the data clusters is too small.

Document Clustering: Document (text) clustering is of interest to a wide range of professions. These include the legal, medical and intelligence communities. Boolean based search and retrieval methods have proven inadequate when faced with the rigors of the current production volume of textual material. Furthermore, Boolean searches do not capture conceptual information.

A suggested approach to the problem has been to somehow extract conceptual information in a manner that is amenable to numeric analysis. One such method is the coding of a document as a collection of trigrams and their frequency of occurrence recorded. A trigram is a collection of any three characters, such as AFV, KLF, OID, etc. There are therefore 26^3 trigrams. White space and punctuation are not included. A document can then be represented as segmented into a specific set of trigrams

starting from the beginning of the text streaming from that document. The resulting set of trigrams from that document and their frequencies are characteristic. If documents in a set have similar trigram sets and frequencies, it is likely that they concern the same topic. This is particularly true if only specific subset of trigrams are examined and counted. The question is, which set of trigrams are descriptive of any concept. A learning algorithm according to the invention can answer that question.

A corpus of 100 English language documents from the Financial Times, The Art of War and the collection of presidential State of the Union addresses was compiled. The corpus was randomly segmented into training and testing corpi. All documents were assigned a value of either 0 or 1, where 0 indicated undesirable and 1 indicated desirable. The learning algorithm searched through the trigram set and identified a set of trigrams that separated the two classes of documents. The resultant model was in 25 dimensions with the decision boundary set at 0.35 the maximal distance allowed in the space. The classifying algorithm utilizes only 25 of the possible 17,576 trigrams. On testing the results in the table obtained.

	Actual Classification 0	1	Totals
Assigned Classification 0	22	2	24
1	6	24	30
Totals	28	26	54

Table: A Confusion Matrix. Actual values are read vertically and the results of an algorithm according to the invention are read horizontally.

The results show that algorithm correctly identified 24 of the 26 documents that were of interest and correctly screened out or rejected 22 of the 26 documents that were not of interest.

16

Evaluation of Biological States: The above-described learning algorithm was employed to develop a classification for prostatic cancer using SELDI-TOF mass spectra (MS) of 55 patient serum samples, 30 having biopsy diagnosed prostatic cancer and prostatic serum antigen (PSA) levels greater than 4.0 ng/ml and 25 normals having PSA levels below 1 ng/ml. The MS data was abstracted by selection of 7 molecular weight values.

A cluster map that assigned each vector in the training data set to a homogeneous data cluster was generated. The cluster map contained 34 clusters, 17 benign and 17 malignant. Table 1 shows the location of each of data cluster of the map and the number of samples of the training set assigned to each cluster.

The classifying algorithm was tested using 231 samples that were excluded from the training data set. Six sets of samples from patients with various clinical and pathological diagnoses were used. The clinical and pathological description and the algorithm results were as follows: 1) 24 patients with PSA >4 ng/ml and biopsy proven cancer, 22 map to diseased data clusters, 2 map to no cluster; 2) 6 normal, all map to healthy clusters; 3) 39 with benign prostatic hypertrophy (BPH) or prostatitis and PSA < 4 ng/ml, 7 map to diseased data clusters, none to healthy data clusters and 32 to no data cluster; 4) 139 with BPH or prostatitis and PSA >4 and <10 ng/ml, 42 map to diseased data clusters, 2 to healthy data clusters and 95 to no data cluster; 5) 19 with BPH or prostatitis and PSA > 10 ng/ml, 9 map to diseased data clusters none to healthy and 10 to no data cluster. A sixth set of data was developed by taking pre- and post-prostatectomy samples from patients having biopsy proven carcinoma and PSA > 10 ng/ml. As expected each of the 7 pre-surgical samples was assigned to a diseased data set. However, none of the sample taken 6 weeks post surgery, at a time when the PSA levels had fallen to below 1 ng/ml were not assignable to any data set.

17

When evaluating the results of the foregoing test, it should be recalled that the rate of occult carcinoma in patients having PSA of 4-10 ng/ml and benign biopsy diagnosis is about 30%. Thus, the finding that between 18% and 47% of the patients with elevated PSA, but no tissue diagnosis of cancer, is consistent with that correctly predicts the presence of carcinoma.

IN THE CLAIMS

I claim:

1. A method of classifying Objects using a vector space having multiple preclassified data clusters comprising the steps of :

   a. inputting a data stream that describes the Object;

   b. abstracting the data stream to calculate an Object vector that characterizes the data stream;

   c. identifying the data cluster, if any, within which the Object vector rests;

   d. assigning to the Object the status of the identified data cluster or, if no cluster is identified, assigning to the Object the status of atypical.

2. The method of claim 1, wherein abstracting is performed by a process comprising selecting between 5 and 25 data points from the data stream.

3. The method of claim 1, wherein identifying is performed by a process comprising computing the Euclidean distance between the centroid of a data cluster and the Object vector.

4. The method of claim 1, wherein identifying is performed by a process comprising computing the normalized vector product of the Object vector and representing the centroid of a data cluster.

5. The method of claim 1, wherein each data cluster is preclassified as having one of

19

two status conditions.

6. The method of claim 1, wherein each data cluster is preclassifed as having one of three status conditions.

7. The method of claim 1, wherein the data streams consist of between 1,000 and 20,000 data points.

8. The method of claim 1, wherein the length of the data streams consist of at least 1,000 data points.

9. A method of constructing a classifying algorithm by using a set of preclassified Objects, each Object being associated with a data stream, where the algorithm is characterized as having multiple data clusters of predetermined extent in a vector space of a fixed number of dimensions, comprising the steps of:

a. providing the set of the data streams associated with the preclassified Objects;

b. selecting an initial set of logical chromosomes that specify the location of a predetermine number of points of the data stream;

c. calculating an Object vector for each member of the set of data streams using each chromosome;

d. determining a fitness of each chromosome by finding the locations in the vector space of a multiplicity of non-overlapping data clusters of predetermined extent that maximize the number of Object vectors that rest in data clusters that contain only identically classified Object vectors, wherein the larger the number of such vectors the larger the fitness of the logical

20

chromosome;

e. optimizing the set of logical chromosomes by an iterative process comprising reiteration of steps (c) and (d), terminating logical chromosomes with low fitness, replicating logical chromosomes of high fitness, recombination and random modification of the chromosomes;

f. terminating the iterative process and selecting a logical chromosome that allows for a optimally homogeneous set of non-overlapping data clusters, wherein the attributive status of each cluster of the optimally homogenous set is the classification of the Object vectors that rest within the data cluster; and

g. constructing a classifying algorithm that classifies an unknown Object by a process comprising calculating an unknown Object vector using the selected logical chromosome and classifying the unknown Object according to the attributive status of the data cluster of the optimally homogenous set of non-overlapping data clusters in which the unknown Object vector rests.

10. The method of claim 9, wherein the fixed number of dimensions is between 5 and 25.

11. The method of claim 9, wherein the number of preclassified Objects is between 20 and 200.

12. The method of claim 9, wherein the initial set of logical chromosomes is randomly selected.

21

13. The method of claim 9, wherein the initial set of logical chromosomes consists of between 100 and 2,000 logical chromosomes.

14. The method of claim 9, wherein the extent of each data cluster is equal.

15. The method of claim 9, wherein the extent of each data cluster is determined by a Euclidean metric.

16. The method of claim 15, wherein the extent of each data cluster in a dimension is a pre-determined fraction of the range of the Object vectors in the dimension.

17. The method of claim 9, wherein the metric that determines the extent of each data cluster is a function of a fuzzy AND match parameter with a vector characteristic of the data cluster.

18. The method of claim 9, wherein the location of each data cluster of the optimally homogenous set is the centroid of the Object vectors of preclassified Objects that rest in the data cluster.

19. The method of claim 9, wherein the location of each data cluster of the optimally homogenous set is the centroid of the Object vectors of preclassified Objects that rest in the data cluster.

20. The method of claim 9, wherein the location of each data cluster of the optimally homogenous set is the centroid of the Object vectors of preclassified Objects that rest in the data cluster.

21. A software product for a general purpose digital computer, accompanied by instructions that the product can be used to perform the method of claim 1 or of claim 9.

22

22. A software product, which performs or causes to be performed on a general purpose digital computer the method of claim 1 or claim 9.

23. A general purpose digital computer, programmed so as to performs or cause to be performed the method of claim 1 or claim 9.

INTERNATIONAL SEARCH REPORT

	Ii Ional Application No
	PCT/US 01/19376

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06K9/62

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
WPI Data, IBM-TDB, PAJ, EPO-Internal, COMPENDEX, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 93 05478 A (BECTON DICKINSON CO) 18 March 1993 (1993-03-18) page 10, paragraph 2 -page 14, last paragraph; figure 1	1-3,5-8
Y		9-23
	-/--	

[X] Further documents are listed in the continuation of box C. [X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
1 November 2001	09/11/2001

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Granger, B

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 01/19376

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
Y	KIEM H ET AL: "Using rough genetic and Kohonen's neural network for conceptual cluster discovery in data mining" NEW DIRECTIONS IN ROUGH SETS, DATA MINING, AND GRANULAR-SOFT COMPUTING. 7TH INTERNATIONAL WORKSHOP, RSFDGRC'99. PROCEEDINGS (LECTURE NOTES IN ARTIFICIAL INTELLIGENCE VOL.1711), NEW DIRECTIONS IN ROUGH SETS, DATA MINING, AND GRANULAR-SOFT COMPUTING. 7, pages 448-452, XP001034347 1999, Berlin, Germany, Springer-Verlag, Germany ISBN: 3-540-66645-1 the whole document	9-23
A	CHANG E I ET AL: "USING GENETIC ALGORITHMS TO SELECT AND CREATE FEATURES FOR PATTERN CLASSIFICATION" INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IJCNN). SAN DIEGO, JUNE 17 - 21, 1990, NEW YORK, IEEE, US, vol. 3, 17 June 1990 (1990-06-17), pages 747-752, XP000146645 the whole document	9-23

INTERNATIONAL SEARCH REPORT

nformation on patent family members

In tional Application No
PCT/US 01/19376

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9305478 A	18-03-1993	AT 151546 T	15-04-1997
		DE 69218912 D1	15-05-1997
		DE 69218912 T2	09-10-1997
		EP 0554447 A1	11-08-1993
		ES 2102518 T3	01-08-1997
		JP 2581514 B2	12-02-1997
		JP 6501106 T	27-01-1994
		US 5627040 A	06-05-1997
		WO 9305478 A1	18-03-1993
		US 5739000 A	14-04-1998
		US 5776709 A	07-07-1998
		US 5795727 A	18-08-1998